FIT5202

Data Processing for Big Data

Project Proposal

Group 01

Wongnaret Khantuwan, Rex (29289440)
Thi Bich Ngoc Hoang, Rita (28496337)
That Bao Thu Ton, Thomas (28576322)
Lam Wai Chun, Tommy (28422635)

OCTOBER 29

Monash University

MONASH University

1. Business requirement

An investment banking is considering to invest in three areas – streaming data, static data and hardware. Based on the given information of the bank, our group is going to propose a project investigating in three main area – credit card transactions, transportation investment, and hardware investment.

Streaming data - credit card transaction data

According to a report made by Australia Post, Australian spent around $21.3 billion purchasing online in 2017. Nearly 1 in 10 items is bought online (aupost, 2018). With a massive amount of data from credit card transactions, the bank wants to monitor the payments made by their customers against any misuses of their cards.

The project proposes an improvement in the hardware system to ensure that the bank can receive and analyse customers' credit card transaction record in the real-time basics with the rate of 1500 files/minutes (each file is about 100KB in size). Every data file must remain available for within 24 hours. The cluster should be able to process streamed data within the interval of 3 seconds and be able to complete the largest batch job in one day (24 hours).

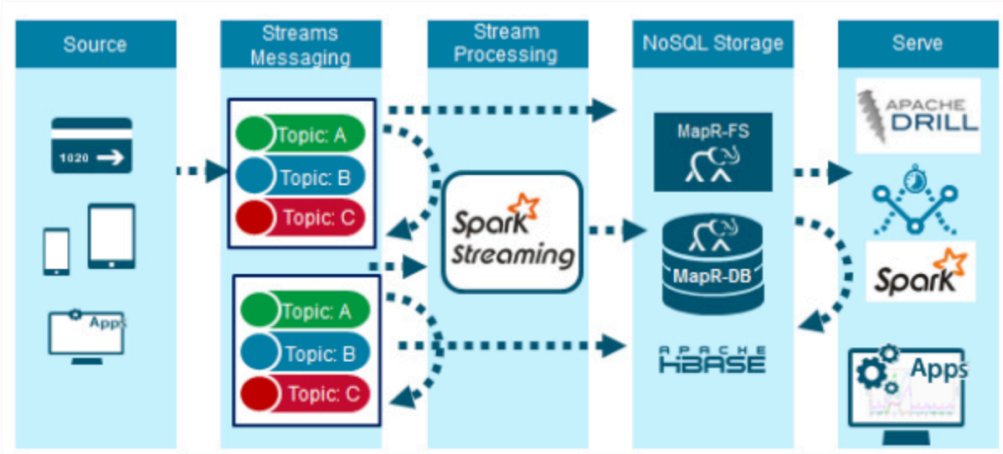


Figure 1: Example of processing credit card transaction data (McDonald, 2016)

Static data - transportation investment

Airlines industry is a promising field. In 2017, 849.3 million passengers were carried by the US airlines systems, surpassing the previous number in 2016 by 3% (Bureau of Transportation Statistics, 2018). Understanding the profitability of this industry, the bank is preparing to expand the market share of airport ownership in the US market. Hereafter, they require extensive analyses on external data source available on the web by the United States Department of Transportation – Bureau of Transportation Statistics (BTS) to support their investment decisions. This data source provides insights regarding the number of passengers, the number of flight or the situation of specific routes, hence, tells the bank which airport they should spend money in.

The bank's demand for data analysis on the external data source and historical transactions lead to another requirement for this project regarding cloud storage for static data. To update data appropriately, the cloud storage must be able to store 100,000 datasets with the size of 150 MB each at any given time.

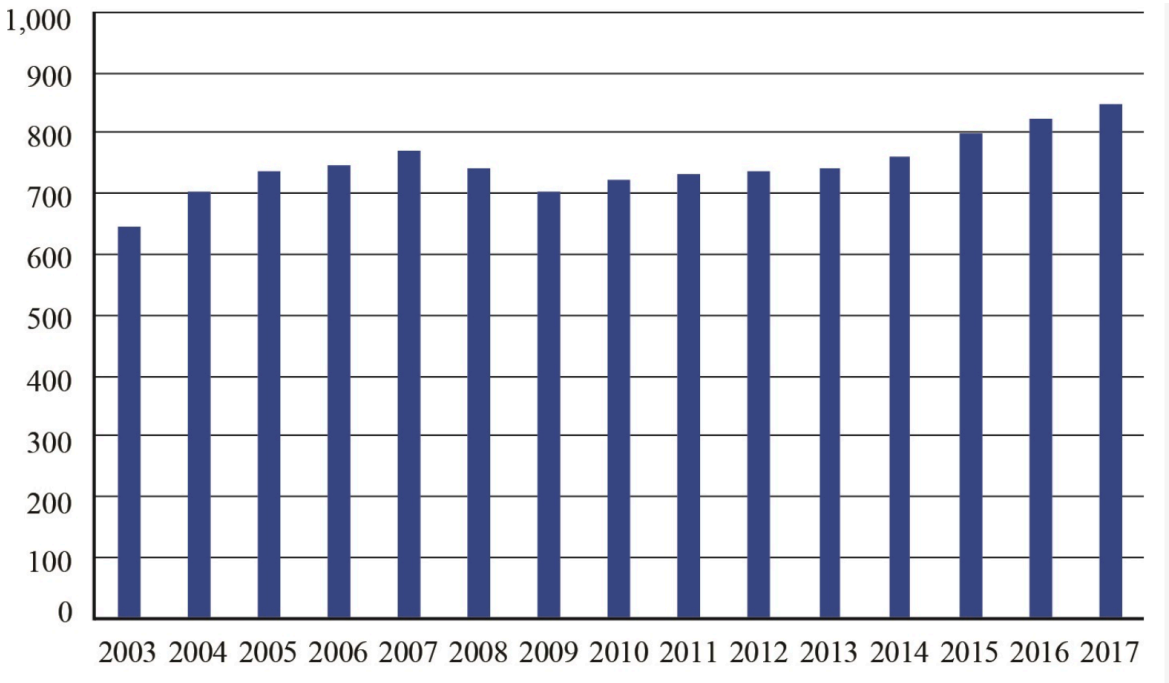


Figure 2: Annual Passengers on All U.S. Scheduled Airlines (Domestic & International), 2003-2017 (Bureau of Transportation Statistics, 2018)

Computer hardware

Generally speaking, large enterprises like banks often encounter complicated issues concerning the availability of the data and performance of the high-traffic website and applications across divisions (rackspace). Cloud computing, as shared pools of configurable computer system resources and high-level services, can help the bank increase their operational efficiency, productivity and agility by relying on shared resources. We will weigh the advantages and disadvantages of each hardware and software option to improve the bank's investment decisions.

The choice of computer hardware must able to meet requirements for processing streaming data and static data. Also, the design of the system will ensure the scalability and allow integration of external source.

2. Business case

This enterprise focuses on maintaining the bank's activities and services like online payment services. Besides, the bank also tries to protect customers from scams and frauds. Finally, the bank aims at maximising profit by utilising their current data resources and infrastructure as well invests in other fields. From the requirement and expectation of this investment banking, our group is going to list all related use cases and analyse possible benefits and risk associated with each use case.

Fraud detection

Along with the development in e-commerce, there is increasingly growth in card fraud. According to Westpac bank, card fraud happens when someone steals credit card details of another person and uses them over the phone or on the Internet to purchase goods and products (Westpac, 2018). The fraud can be recognised in two ways. First, based on the historical transaction data of the user, the bank sends the notification to request the confirm from this user about an abnormal payment. Abnormal payments can be defined as a payment made overseas or a payment with the amount over a certain limit. Second, the user reports an unauthorised transaction made from their account. The streaming data of online transactions, in this case, is a significant factor to detect and prevent scams. As the project designs a system with delays of three seconds in streaming data, this system will facilitate timely preventative actions from the bank such as temporary lock a particular card to protect the user from the risk of the scam.

This application of the project will involve stakeholders in both the customer side and the bank side. If this project success, customers will be able to enjoy shopping online without worrying about scams while the bank can gain reputation regarding high security. However, during the upgrade period, the bank can take the risk from minor issues with the payment system. Furthermore, frequent notifying about wrongly possible scams from the bank might negatively influence the user experience.

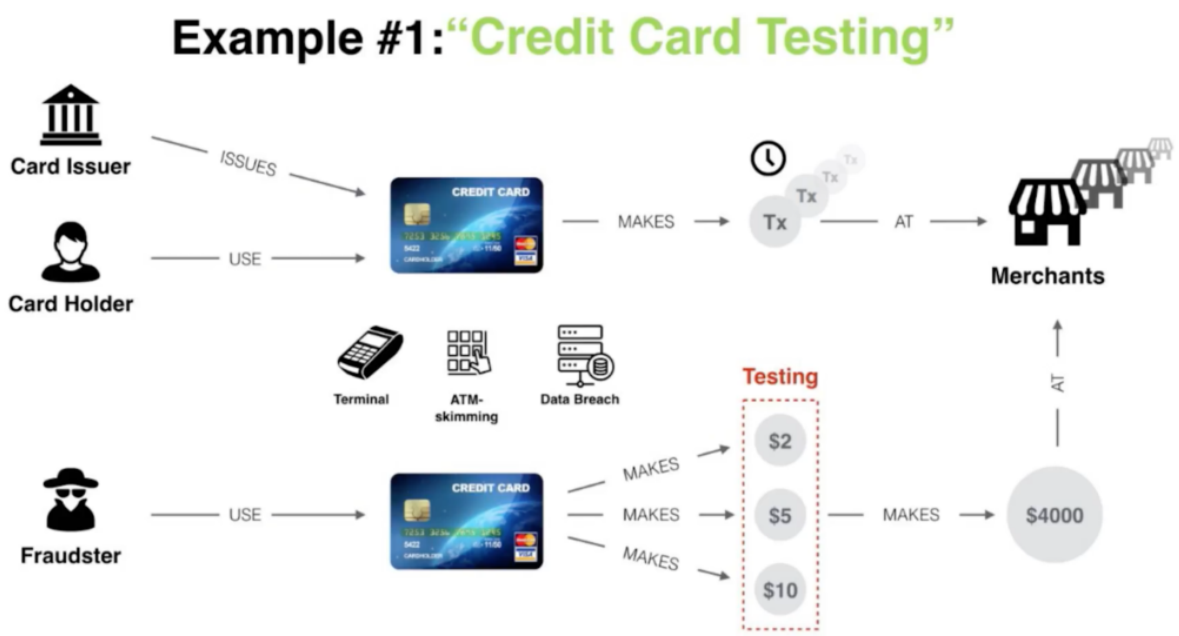


Figure 3: Example of credit card fraud detection (Mathur, 2017)

Investment

This project proposes the use of 15TB cloud storage server, which opens new opportunities regarding the accessibility of data and the integration of external data sources. In other words, the banking officials can have the most updated insights for any given industries by querying data from this new system. For example, the bank can identify which airport was the busiest airport in the US in 2017 from data provided by the Bureau of Transportation Statistic; therefore, spend more money on this airport.

It is undeniable that the analyses on static data can create competitive advantages for the bank when investing in any fields. The stakeholders, in this case, are the bank and data providers. In some circumstances, the bank needs to pay money to get access to some external data sources.

Loyalty program

Once all external data and historical data can be retrieved with ease, the company can deploy different programs to attract more customers and encourage them to pay for services and product invested by this bank. In particular, our group suggests the bank to combine the publicly available data from BTS with their privately-owned data of credit card transaction to customise promotion and loyalty program to attract more customers. For example, the bank might give their loyal credit card users discounts in terms flight tickets, merchant products, and services at their own airports. This strategy will encourage users spending more to become a high-rank customer to enjoy the perks as well as motivate them to travel more to this airport.

In respect of loyalty programs, stakeholders might include the bank, credit card users, merchants in the airport and airlines. The primary challenge in this use case is that the bank needs to negotiate the deal with the suppliers of products and services in this program to gain win-win benefit for both the bank, the customers, and the suppliers.

Figure 4: Example of relationship between big data and loyalty program in banking (Andreadakis, 2013)

3. Design Analyses

Data Sources

Fraud Detection

This task will require two sets of data, in order to achieve the fraud detection whenever there is any abnormal transaction happened.

First of all, in order to build a model to detect any potential fraud over phone or internet transactions, historical transaction data will be required including the transaction amounts, bank information, credit card information. All these data might include certain level of privacy, they can only be provided by our client, bank enterprise.

After the model built, the next step is to recognise the potential fraud on retail banking including credit cards for online purchase, and online transactions. some of the transactions are within local regions while some of them involve oversea transfer. All these transactions data have to be instant, since we may need to feedback our customers if there are suspicious payment. Therefore, this data will be streaming data.

Investment in transport infrastructure

For the investment proposal in transport infrastructure, we are expecting to have a deeper understanding on aviation industry in order to predict the profitability of expanding the market share. Therefore, some data on aviation industry will be required to assist the analysis. The open data by BTS is one of the key data adopted in this project. This is an independent statistical agency, aiming to supply reliable and accurate information about transportation. They integrate the data from a broad range of sources, including information on flow of traffic and passenger, aviation performance, employment and financial condition, etc. (BTS, 2018) This data is comprehensive for the investigation on transportation investment.

In order to have a comprehensive understanding on aviation industry, Skytrax data will also be used. This is a cleaned dataset, collaborating all the reviews found on Skytrax, an Air Travel Review website from independent customer forum. With more than 40k reviews from customers on Airline, Airport, Seat and Lounge. (Nguyen, 2015) It is strongly believed that this dataset can assist with BTS data, on analysing the potential profitability

of aviation industry and providing timely, accurate and comprehensive information to the bank.

Data Ingestion

Data ingestion is the process to access and import all the data into the database for usage or storage. All the data will be prioritized and categorized in order to facilitate the data flow for later parts of data handling (imanuel, n.d.).

Depends on the type and source of the data, data ingestion can be performed in different ways:

For the static data in Investment model and Fraud Detection, as they are not real-time data, they will be loaded in batches. This means the data can be imported in discrete chunks of data at intervals; while for the streaming data on banking transactions, real-time data ingestion will be required.

Tools will be required to prioritize all these data, validate through all the files and hence dispatch them to the correct destination in order to maximize the ingestion efficiency.

Therefore, for the data ingestion, we will adopt the **Apache Kafka**, which is an open-source message broker with strong durability and fault-tolerance (imanuel, n.d.).

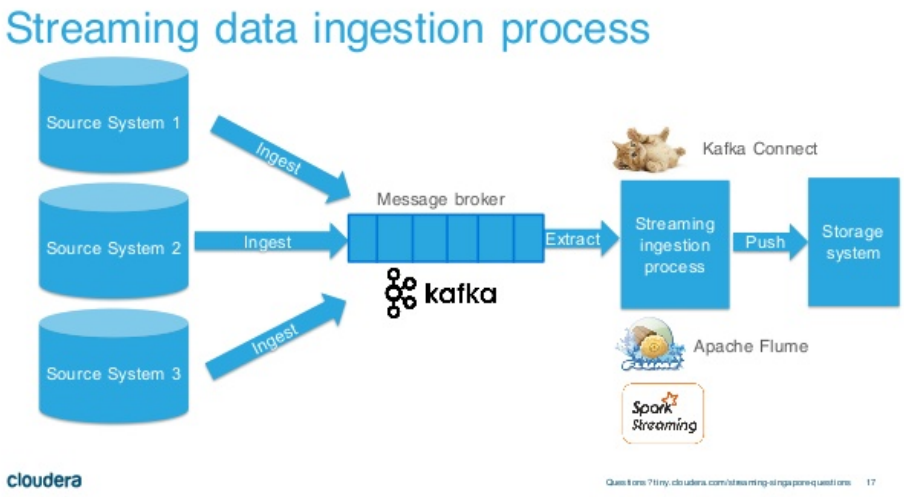


Figure 5: Example of data ingestion on streaming data with Apache Kafka

Data storage design considerations

In this project consist of 2 categories of data streaming data and static data. For capability to store a large amount of data that might need the storage at least 15 TB in total, and also process and serve the data on time. The most suitable architecture of our computer cluster for this project is based on the Hadoop and run on top with the Spark system. With this architecture, the data will be distributed into each data node in the cluster for faster parallel processing, and Hadoop is providing build-in false tolerance mechanism (data-flair support., 2016) for protecting the data in the system. Future more, Hadoop cluster is also supporting scaling ability (Appuswam, Gkantsidis, Narayanan, & Hodson, 2013) for expanding this cluster in the future in case of part upgrade or the more machines are needed.

Cluster design

Since the cluster is based on Hadoop as mentioned above. First, this cluster is initial with 20 machines, and we will split into two groups of machines as shown in the following diagram:

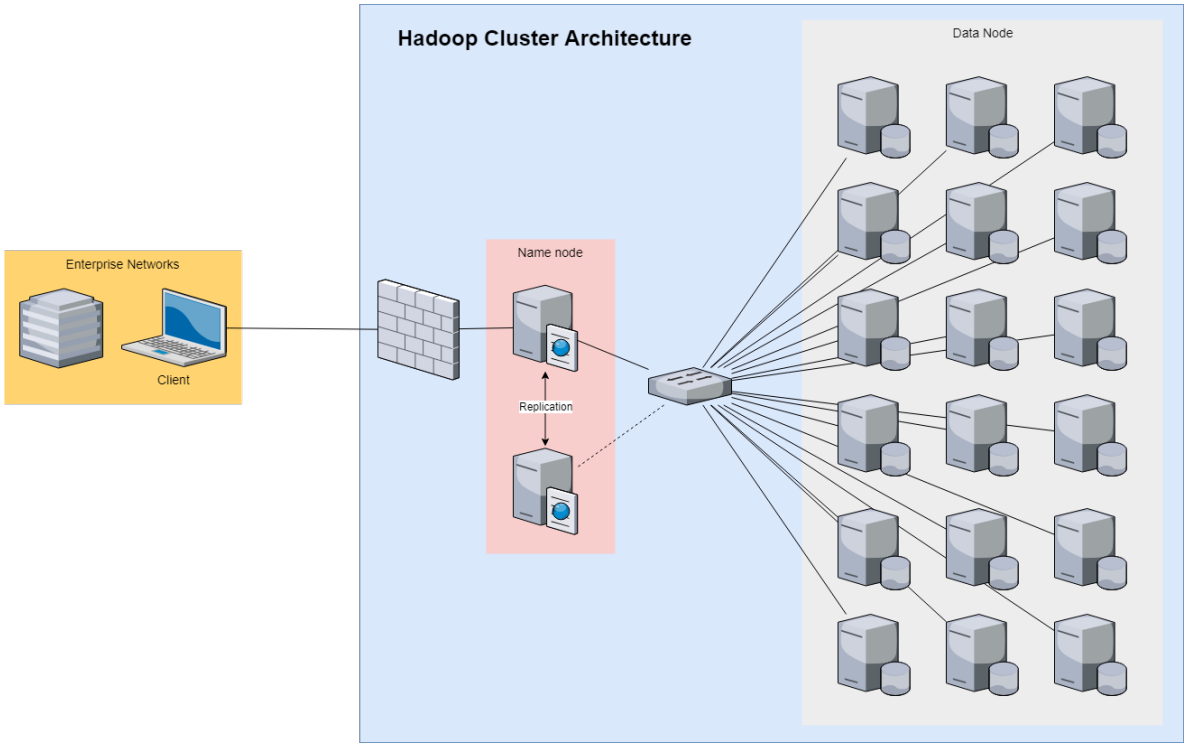


Figure 6: Hadoop cluster Architecture

According to the cluster architecture diagram:

- 2 name node machines: for safety, two machines are assigned to be the name node. One of them will be an online machine, and another one is the replicate server for backing up the first one in case of crashing.
- 18 data node machines: All the data for this project will be distributed to these nodes, and all the processing task will happen on these machines.

In term of the storage, the consideration is looking onto two aspects.

- Physical: Each machine in the cluster consist of 2 types of disk:
    - Solid State Drive (SSD): each machine is installed with an SSD for the OS and cache. This can help the machine is a faster response.

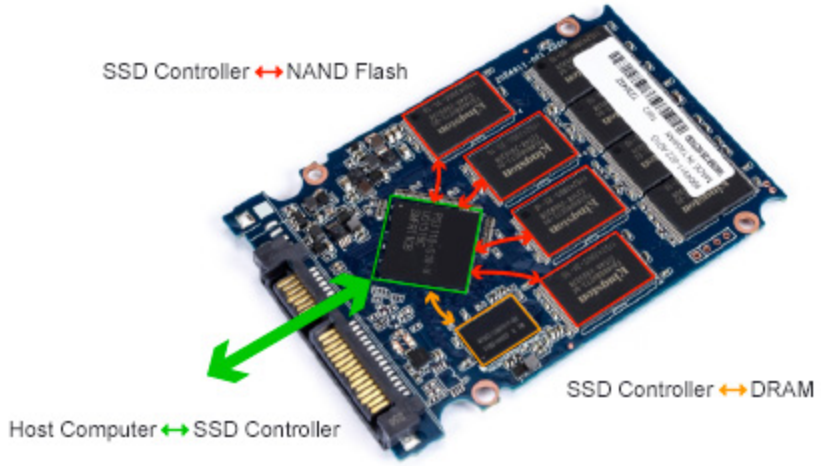


Figure 7: Solid-state drive. **(Kingston Technology Company, n.d.)**

    - An ordinary hard disk drive: that installed in the machine, this will be the main storage for data HDFS data storage. Every machine will install with a couple of the HDD.

Figure 8: Harddisk drive. (Wikipedia, n.d.)

- Logical; Based on the Hadoop and Spark framework, this made the cluster have the logical storage as:
    - Hadoop-DFS: the basic file system for the distributed data framework
    - Spark data Parquet: for the faster computing on with the Spark framework; The Spark Parquet will be stored in the HDFS file system, which is some benefits such as less disk IO by read/write a large amount of data in one time and Spark SQL is run faster with Parquet (Chen, 2016).The following figure shows how Spark working with the data and Parquet in the HDFS file system.

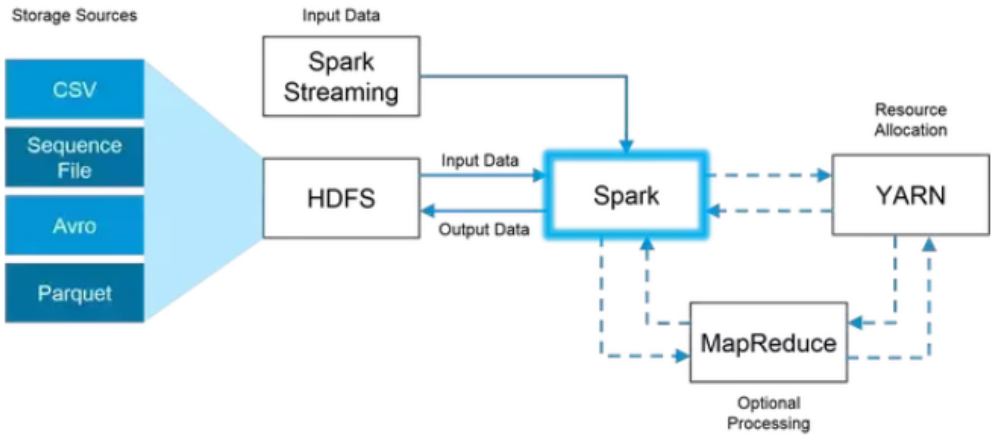


Figure 9: The architecture of Spark and Hadoop file system (Dayananda, 2017)

Data processing considerations

According to the Business case and data source, this project is focused on analysing the credit card transaction data and Airline transportation data. Thus, the data processing for this project need to be considered in many areas:

Fraud detection

The fraud detection task for credit card transaction, a lot of machine learning algorithms has proposed for detecting fraudulent activity such as Bayesian network, Support vector machine, etc. (Awoyemi , Adetunmbi, & Oluwadare, 2017; McDonald, 2016) According to this reason, the SparkML is needed for data processing. The example of building a solution for credit card fraud detection consists of 2 phases as the following figure.

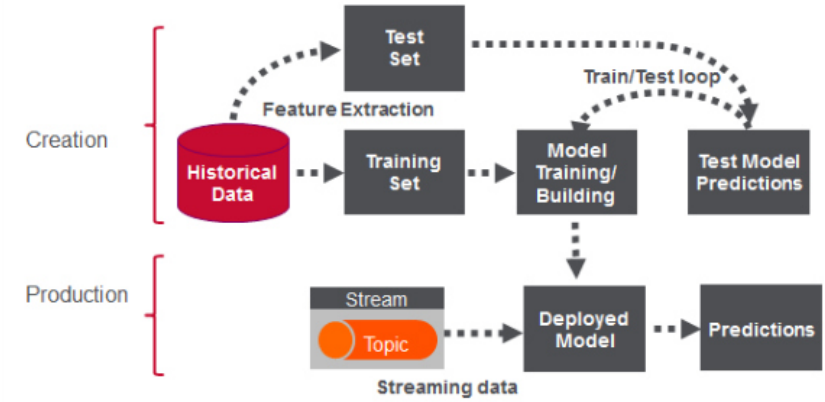


Figure 10: Two phases of building a fraud detection system (McDonald, 2016)

The first phase involves analysis of historical data to build the ML model. After that, the model in production is used to make predictions on live events.

Loyalty program

The main idea for the Loyalty program generating algorithm is based on the question "Which flight/Airline is the most suitable for gaining the number of credit card customer." (Mercator Research, 2013) Since the Airline transportation data is a graph based dataset. The graph algorithm with Spark GraphX is suitable algorithm for analysis the flight data. For example, the shortest path algorithm can be used for flight planning in term of time and cost; the maximum flow algorithm can be used for calculating the number for a capacity of passengers for each route that can help the company can design the loyalty program based on the number of passengers.

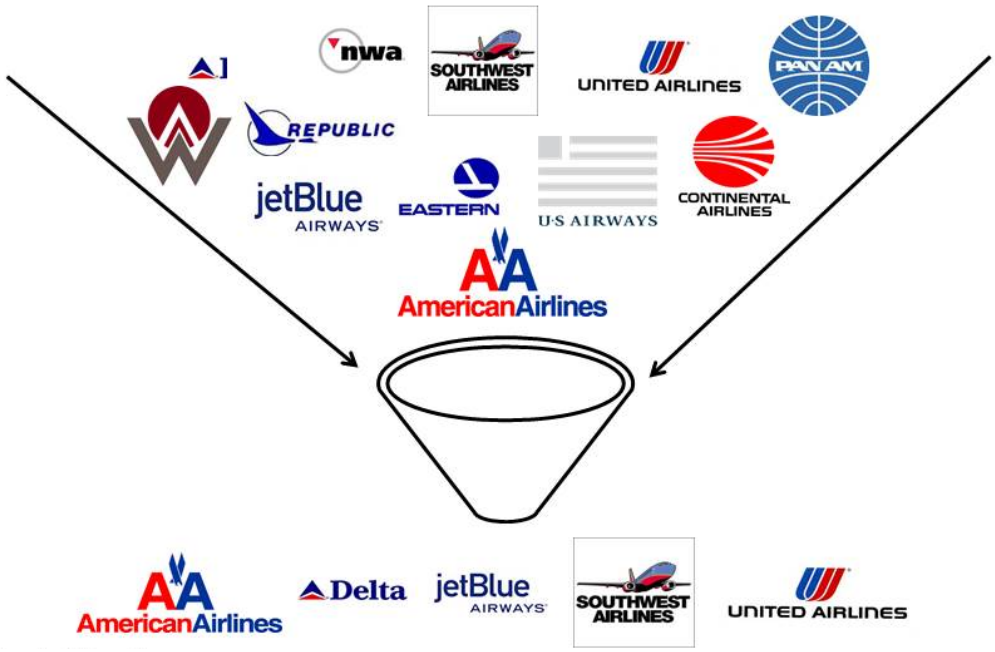


Figure 11: The algorithm suggests the airline company as the candidate for the loyalty program. (Mercator Research, 2013)

Furthermore, since the credit card always update. The batch processing and stream processing is required for updating the data along with SparkSQL. And this project is implemented in Spark framework, we can use Spark Streaming for receiving live input credit card transaction streams, divides the data into batches, then processed by the Spark engine to generate the final stream of results in batches as shown in the diagram.

Spark Streaming (SPARK, 2018)

Data output provisions:

This project can return the result into a number of output such as:

- **Data mining:** Since this project exploring the insight data for Airlines transportation and the credit card transactions using data mining principles. The result will be suggest how to choose the flight/Airline for creating the royalty program that can help the airline transportation and credit card business growing by gaining the number of the credit card used in airline transportation market.

- **Machine Learning:** The credit card transaction data will be processed for detecting fraud using the SparkML machine learning algorithm. The result of this process will be the indication note for the specialist for making sure and following the legal business process.

- **Online queries:** This project also provides business intelligence for easier understating and help business strategy planning. The output will be displayed as a dashboard and support online query for the related question in the future. Thus, the business intelligence system (BI) is useful for this task.

4. Costing

Static Data

In this case of the static data, we can use cloud technology to store to reduce the overhead cost of the business such as electricity, maintenance and replacement fee. Although the most common concern of cloud storage is the security, our static data can be stored on the cloud storage because the static data is from the external data source. Hence, it is safe to store this type of data on the cloud.

According to the business requirement, the cloud server must be able to store 100,000 CSV files, each of the file is 150MB in size. So, it should be able to save 15TB at any given time.

Amazon Web Service (AWS)

x1e.2xlarge - High performance databases, in-memory databases

- High frequency Intel Xeon E7-8880 v3 (Haswell) processors
- 244GB Ram
- 240 GB SSD

$2.418 USD per Hour per instance => **$21,181.68 USD/instance/year (24/7 running)**

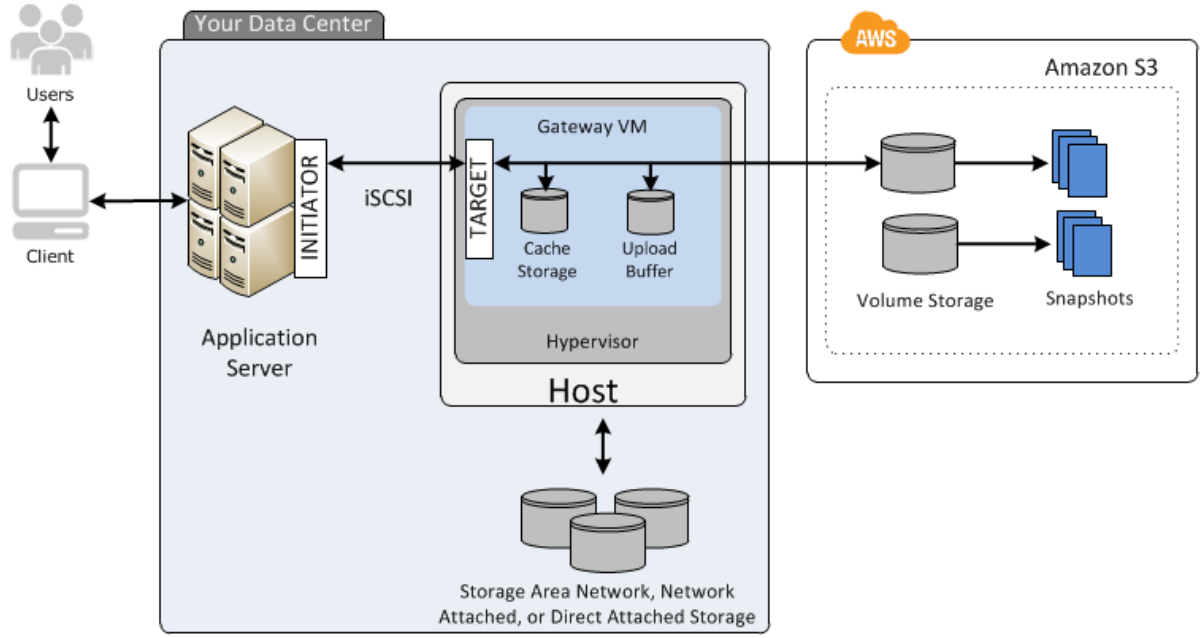


Figure 13: AWS cloud storage architecture. (google cloud, 2016)

Google Cloud

Google cloud offers computing architecture for both dynamic and static data. However, due to security issues, we only need the solution for cloud storage for our static data (see Figure 14).

- CPUs :16
- 104 GB Ram
- 375 GB SSD

$890.67 USD per 1 month per instance => **10,688.04 USD/instance/year (24/7 running)**

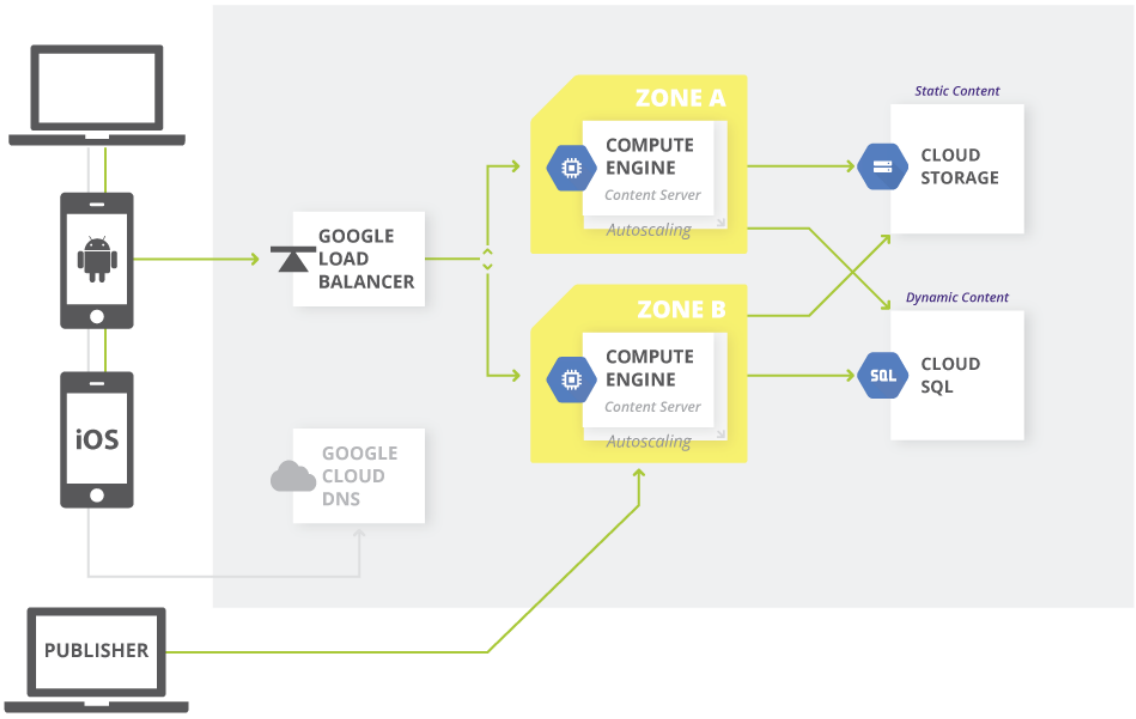


Figure 14: Google Cloud Platform cloud storage architecture.
Retrieved from (google cloud, 2016)

Dynamic Data

- The transaction data of the bank requires a high-security solution. Cloud is cheaper, but security is the big issue as the server may not be placed within the country or on-site. Hence, building our server is considered more expensive in term of the maintenance fee, complex installation and deployment, long testing time but more secured for transactions data.

Hardware Costing:

Physical Server:

Figure 15: PowerEdge R330 rack server. (DELL, 2018)

PowerEdge R330 rack server = 1 RU
Intel Xeon E3-1230 v6 3.5GHz, 8M cache, 4C/8T, turbo (72W)
8x 16GB UDIMM, 2400MT/s, Dual Rank = 128GB
80GB SSD SATA Mix Use 6Gbps 512 2.5in Hot-plug AG Drive,3.5in HYB CARR
4x 1TB 7.2K RPM SATA 6Gbps 3.5in Hot-plug Hard Drive
On-Board LOM 1GBE Dual Port
Price 13,620.50 USD

Our rack server has 128Gb ram to handle the streaming data. We designed the streaming data batch will be loaded and processed on ram (in-memory) to enhance processing performance on the HDFS.

For storage, the Physical Hard Disk Drive for one server unit is designed to have the capacity of 4TB to store the raw transaction datafiles for after streaming analytics. We also need one of the Solid-State Drive (SDD) with the Windows Server Installed on each rack server to improve the Operating System boosting time of the server.

For further expansion. We will need from 18-20 machines that will cost:
PowerEdge R330 rack server X 20 = 272,410 USD

Furthermore, we also need overhead cost (maintenance, electricity, etc.) to operate the service. For example, a PowerEdge rack consumes 7,446 kWh per year (Maddox, 2013) Suppose our server is placed in Australia which has the electricity price at 25cents/kWh (Mountain, 2012). Hence, in one year, 1,861.50 AUD ≈ 1,400 USD per year for electricity to operate **one machine**. For 20 machines, it will cost about 28,000 USD in electricity cost to run the total rack of 20 nodes.

Server Rack:
APC AR3104 24U Server Racks/Cabinets X1 = $1054 USD

Figure 16: APC AR3104 24U Server Racks/Cabinets. (NEWEGG, 2018)

We need a server rack to store our machines, so we design to purchase a rack storage capacity up to 24 servers for further expansions.

Software Costing:

OS: Red Hat Enterprise Linux Server: $349 USD/node

Server storage: HDFS

Database Design: Hadoop cluster

Data analytics software: Apache Spark

Bibliography

2017 Annual and December U.S. Airline Traffic Data. (2018, January). Retrieved October 2018, from bts.dot.gov: https://www.bts.dot.gov/newsroom/2017-annual-and-december-us-airline-traffic-data

rackspace. (n.d.). *cloud-computing-advantages.* Retrieved October 2018, from www.rackspace.com: https://www.rackspace.com/en-au/library/cloud-computing-advantages

McDonald, C. (2016, May). *Real Time Credit Card Fraud Detection with Apache Spark and Event Streaming.* Retrieved from mapr.com: https://mapr.com/blog/real-time-credit-card-fraud-detection-apache-spark-and-event-streaming/

google cloud. (2016, October). *Architecture: Content Management*. Retrieved October 2018, from cloud.google.com: https://cloud.google.com/solutions/architecture/contentmanagement

DELL. (2018). *DELL.COM*. Retrieved from www.dell.com: https://www.dell.com/en-us/work/shop/povw/poweredge-r330

NEWEGG. (2018). *NEWEGG*. Retrieved from www.newegg.com: https://www.newegg.com/Product/Product.aspx?Item=N82E16816225057

CISCO. (2018). *CISCO*. Retrieved from www.cisco.com: https://www.cisco.com/c/en/us/support/switches/sg100-24-24-port-gigabit-switch/model.html

SPARK. (2018). *Spark Streaming Programming Guide*. Retrieved from spark.apache.org: https://spark.apache.org/docs/latest/streaming-programming-guide.html

Mercator Research. (2013, APRIL 13). *Credit Card Reward Programs: Balancing Loyalty and Profitability*. Retrieved from www.mercatoradvisorygroup.com: https://www.mercatoradvisorygroup.com/Reports/Credit-Card-Reward-Programs--Balancing-Loyalty-and-Profitability/

BTS. (2018, JULY 25). *About BTS*. Retrieved from www.bts.gov: https://www.bts.gov/about-BTS

Nguyen, Q. (2015, Aug 2). *Skytrax dataset and information*. Retrieved from github.com: https://github.com/quankiquanki/skytrax-reviews-dataset

imanuel. (n.d.). *Top Data Ingestion Tools*. Retrieved from www.predictiveanalyticstoday.com: https://www.predictiveanalyticstoday.com/data-ingestion-tools/

Kingston Technology Company. (n.d.). *Data Transfers Within an SSD*. Retrieved from www.kingston.com: https://www.kingston.com/en/ssd/data-protection

Wikipedia. (n.d.). *Hard disk drive*. Retrieved from en.wikipedia.org: https://en.wikipedia.org/wiki/Hard_disk_drive#/media/File:Laptop-hard-drive-exposed.jpg)

Dayananda, S. (2017, Sep 19). *In which scenario would you use Hadoop over Spark?* Retrieved from www.quora.com: https://www.quora.com/In-which-scenario-would-you-use-Hadoop-over-Spark

aupost. (2018). *2018-ecommerce-industry-paper-inside-australian-online-shopping.* Retrieved October 2018, from auspost.com: https://auspost.com.au/content/dam/auspost_corp/media/documents/2018-ecommerce-industry-paper-inside-australian-online-shopping.pdf

Bureau of Transportation Statistics. (2018, January). *2017 Annual and December U.S. Airline Traffic Data.* Retrieved October 2018, from bts.dot.gov: https://www.bts.dot.gov/newsroom/2017-annual-and-december-us-airline-traffic-data

Westpac. (2018). *Credit card fraud.* Retrieved October 2018, from www.westpac.com.a: https://www.westpac.com.au/security/fraud-and-scams/credit-card-fraud/

Mathur, N. (2017, April). *Fraud Prevention with Neo4j: A 5-Minute Overview.* Retrieved from neo4j.com: https://neo4j.com/blog/fraud-prevention-neo4j-5-minute-overview/

Andreadakis, D. (2013, October). *CUSTOMER DATA: A KEY TO BANKS' LOYALTY PROGRAM SUCCESS.* Retrieved from blog.kobie.com: http://blog.kobie.com/2013/10/customer-data-a-key-to-banks-loyalty-program-success/

Appuswam, R., Gkantsidis, C., Narayanan, D., & Hodson, O. (2013, Oct). *Scale-up vs Scale-out for Hadoop: Time to rethink?* Retrieved from www.microsoft.com: https://www.microsoft.com/en-us/research/publication/scale-up-vs-scale-out-for-hadoop-time-to-rethink/

Chen, J. F. (2016, Jan 15). *5 Reasons to Choose Parquet for Spark SQL*. Retrieved from developer.ibm.com: https://developer.ibm.com/hadoop/2016/01/14/5-reasons-to-choose-parquet-for-spark-sql/

Awoyemi , J., Adetunmbi, A., & Oluwadare, S. (2017). *Credit card fraud detection using machine learning techniques: A comparative analysis*. Retrieved from doi.org: https://doi.org/10.1109/ICCNI.2017.8123782

data-flair support. (2016, Jun 13). *Understand HDFS Feature – Fault Tolerance*. Retrieved from data-flair.training: https://data-flair.training/blogs/learn-hadoop-hdfs-fault-tolerance/

Appendix A - Meeting minutes

1st Meeting:

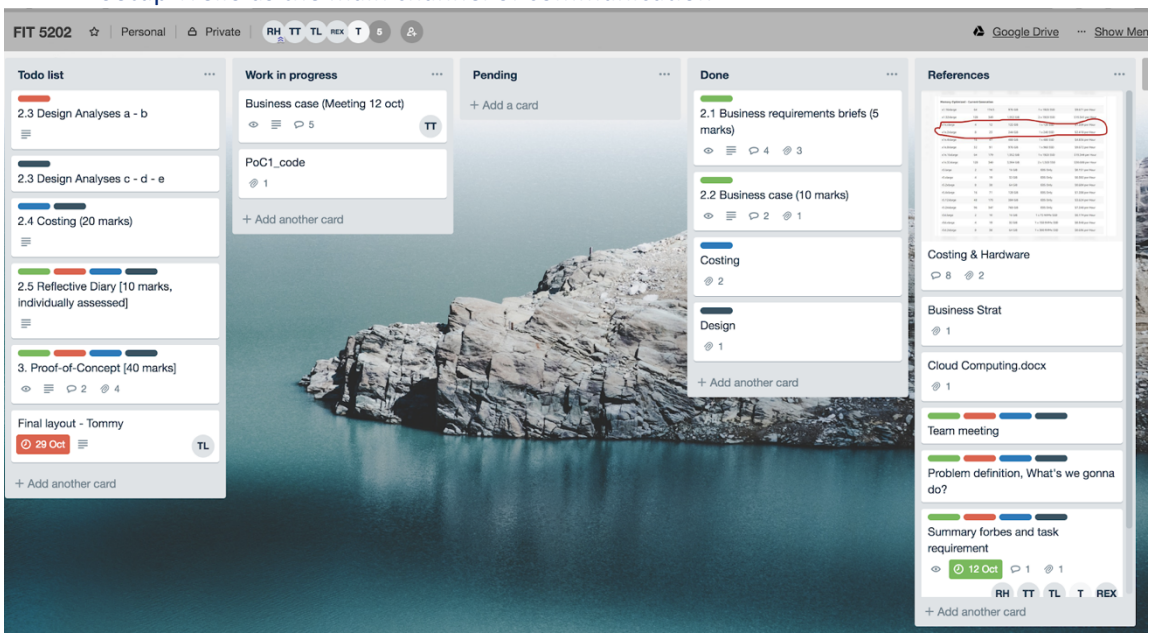
Date: Oct 5, 2018 (Fri)
Time: 14:30 - 17:00
Attendees: Rex, Rita, Thomas, Tommy

MEETING MINUTES

Description:
This is the record for our 1st meeting. In this meeting, we discussed the assessment requirement, set up our Google Drive and Trello dashboard, and allocate the task for each member.

Discussion:

- Go through the assessment
  - What is the business case and requirement?
  - How can we structure our report?
- Clarify some concepts from the lectures and assignment
  - Spark architecture, Database structure (HDFS vs local)
  - Business case
  - Costing
  - Proof of concept
- Setup Trello as the main channel of communication



Task allocation for next meeting:

- 2.1 - Business requirement, 2.2 - Business case: Rita
- 2.3 Design analysis a-b: Tommy
- 2.4 Design analysis c-d-e: Rex
- 2.5 Costing: Thomas

Next Meeting: Oct 9, 2018 (Tue) 20:00

2nd Meeting:

Date: Oct 12, 2018 (Fri)
Time: 15:30 - 16:30
Attendees: Rex, Rita, Thomas, Tommy

MEETING MINUTES

Description

This is the meeting minute for our 2nd meeting. In this meeting, we investigated in-depth about the business requirement and business case. Also, we also tried to handle all the problem of each individual in completing their task.

Discussion

- What is the use case of each area
    - Streaming:
        - Applying fraud detection
        - Unauthorised use of credit cards /stolen cards
        - Abnormal transaction (suddenly pay a large amount of money)
        - Reference: https://www.westpac.com.au/security/fraud-and-scams/credit-card-fraud/ https://mapr.com/blog/real-time-credit-card-fraud-detection-apache-spark-and-event-streaming/ https://www.kaggle.com/mlg-ulb/creditcardfraud

- Airline investment
    - Airline suggestion system (model)
    - Finding insight from the dataset (Which route should we invest in? (Based on number of flights))
    - Combine 2 dataset together (rewards when flight certain routes, loyalty scheme)
    - Reference:https://www.slideshare.net/simplify360/big-data-and-social-media-analytics (Quantas point)

Task allocation for next meeting:

- Finalise all ideas Tuesday (16 Oct)
- Complete 2.1 and 2.2 Business case (Rita) (19 Oct)
- Going to consultation: Rita and Thomas (25 Oct)

Next Meeting: Oct 28, 2018 (Sunday), 14:30

3rd Meeting:

Date: Oct 28, 2018 (Sunday)
Time: 14:30 - 16:30
Attendees: Rex, Rita, Thomas, Tommy

MEETING MINUTES

Description

This is the meeting minute for our 3rd meeting. In this meeting, we combined our work for this assignment and checked the code for POC1 and POC2.

Discussion

- Work integration:
  - Integrate business case and design analysis a,b
  - Align design analysis c,d,e and costing
- Testing and checking the code
  - Checking the result of POC2
  - Discuss implementing POC1
- Writing report and documentation

Task allocation for next meeting:

- Work integration: all members
- Checking bugs of code
  - POC 1: Rex
  - POC 2: Thomas, Rita
- Writing code documentation, drawing flow charts
  - POC1: Rex
  - POC 2: Rita, Thomas
- Designing and final consolidation: Tommy
- Check alignment of all parts: Tommy
- Summarise meeting minutes: Rita

Next Meeting: Oct 29, 2018 (Monday), 15:30

4th Meeting

Date: Oct 29, 2018 (Sunday)
Time: 15:30 - 20:30
Attendees: Rex, Rita, Thomas, Tommy

MEETING MINUTES

Description

This is the meeting minute for our final meeting. In this meeting, we finalised all supporting documents and tested our POC codes.

Discussion

- Finalise the final report (All member)
    - Edit the report template
    - Cut down the word to meet the word count limit
- Summarise the meeting minutes (Rita)
- Check code bugs for POC 1(Rex & Thomas)
- Complete the reflective diary

Our final meeting at Caulfield

Appendix B – Reflective Diary

Thi Bich Ngoc Hoang (Rita) – 28496337

Role: Programmer and Coordinator

Main responsibility

- Write section 2.1 Business requirement and 2.2 Business case
- Implement, draw diagram and document POC2
- Summarise all meeting minutes and supporting documents

Weekly reflective diary

Our project started on 30th September and ended on 29th October. The project lasted for one month. This document will provide my reflective diary about the process of working.

Week 1: 30th September to 6th October

Week 1 was the kicked-off week of our team. In this week, I was responsible for setting up the working environment for our team in Trello. Besides, we also started reading and discussing the assessment. I was confused about the assessment specification, especially regarding the datasets used for this project. Therefore, we decided to ask our tutor and were explained more about the spec. After this week, I learnt how to plan and organise teamwork.

Generally speaking, the environment setting was quite smooth. We also successfully allocated the task for each. For example. I would need to start writing the business case first so that other team members could align their work with mine. The only problem is our lack of understanding of the assessment requirements; however, our tutors helped us to clarify everything.

Week 2: 7th October to 13th October

Our primary focus for this week was brainstorming about the uses cases of three areas stated in the business requirement. As a preparation for the meeting, I summarised all Forbes cases given as reference documents in Moodle and prepared some additional sources of information. Moreover, I created the structure for Section 2.1 Business requirement and Section 2.2 Business case. As we already knew what we are expected for this assignment, I felt everything going well. I learnt how to summarise and analyse ideas from different articles and my teammates.

In the meeting, we discussed different possibilities for use cases such as fraud detection, loyalty program and investment business intelligent dashboard. The good thing is that we all agreed on the business case idea, and I can start writing my part. The sad thing is that we still need to confirm the size of the organisation with our supervisor; however, we cannot match our time to go to the consultation together. Therefore, Thomas and I were assigned to go and check the remaining doubts at the consultation section next week.

Week 3: 14th October to 20th October

There is no meeting this week. We were all under a lot of pressures as this is assignment submission period. All of us were stuck with our assignments, so we decided to work individually. Without any meeting, we focused on communicating and updating our progress via Facebook messenger. I finished my first draft in business requirement and business case on 19th October and uploaded it to Trello and Google Drive so that other team member can align their work with mine. After this week, I learnt how to prioritise the work and be flexible when working in a team.

The good thing is that everything was on the right track regardless we did not have time for face to face meeting. The bad thing is that we did not have enough time to check other people work in this week. We agreed to leave the checking to the next week because we were confident that we still have enough time to complete this assignment.

Week 4: 21th October to 27th October

This week, we speeded up to complete the major parts of this assignment. I finished my final version of the business requirement and business case. Apart from that, I spent the time to program POC2 with the support of Thu in debugging the code. I was a bit rushed because we had less than 10 days left for completing anything. This was a productive week, and I learnt how to deal with graph processing.

The good thing is that we completed the code for POC2 and most parts of the report. However, there was a misunderstanding in POC2 requirement. In the beginning, I use data frame APIs to process data, but the requirement is using the graph to process the data. After knowing this problem, Thomas helped me fixed it together, and we still got the same result. The remaining tasks now were just documentation for POC2 and integrating everything.

Final days: 28th October - 29th October

We now used 100% effort to complete the remaining things, which were integrating all member work and debugging. I draw and documented all the work for POC2 and summarised all of the meeting minutes. Thank for this, I learnt how to comment the code correctly and test the files. Everything was just on time.

We did not meet any problems these days. We stayed together at Caulfield campus to match our work and edit everything. We are a good team, and I am proud of being a part of this.

Conclusion

This project went well. We are a good team and always support each other. If I have a chance to do this project, I hope that I could be their teammate again, and I will try to spend more time to discuss and learn from them.

That Bao Thu Ton (Thomas) - 28576322

Role: Tester + Programmer

Main responsibilities

- Pilot test the PoC1 and PoC2 code.
- Testing bugs, clean codes for PoC1 and PoC2.

Contribution

- Running the code to make sure it works
- Suggest a better approach for the code (fewer lines, cleaner code, more readability for the print out)
- Estimate the costing part
- Proofread the report parts and code documentation.

Activities

Week 1: Starting

In this first week, we sat together to analyse the assignment specifications, then, we find the concepts from lectures and the case studies on the internet that have the similar to our business case. Setting up the Trello our main channel for communication. In the very first start of the project, I contributed my understanding in analyse the business case. Such as how the bank manages fraud of credit card or loan. I have also listed out what will be needed to be done in the coding task

I felt a little bit frustrating because at first, the business case was not so clear. A different team member has different approaches to the case. I was allocated to the task of the cost of both static data and streaming data of the business case. My approach was to base on the previous case studies of the other banking corporation such as the Bank of America, Commonwealth bank in data management. It seems my approach has worked at the first time; the case studies of other banks seem similar to our assignment case study. Different team members have different approaches such as they did not begin with the case study, they began to analyse the system first and then the business case. Our team decided to combine our idea and weighing the ideas which one is more important and come up with the final solution. However, in the first week, the business case had been still not clear for us yet.

Week 2: Brainstorming

This week, we had the second meeting to continue analysing the business case and brainstorming the design, we also tried to help each other with their individual task if they needed. Specifically, we tried to point out the use cases for each area: streaming, static data, airline investment and the data for the proof of concept. This week, I started on working with the proof of concept requirement to align with the teammate who mainly in charge of coding.

However, the result of our analysing of the business case for this week was still unclear and we needed further consultation. So, I and my teammate (Rita) was allocated with the task of asking for consultation for next week.

Week 3: Working independently

We were asked to work independently this week and will have the meeting next week for integrating all of our work. Cross-check each other works and bugs fixing the Proof of concept code. I was in charge of bug fixing, improving the code conciseness and performance. The code of the programmers of the team works well, the outputs are correct. However, it was a little bit off topic when she did not use the graphX library to do the Proof of concept 2 as the assignment requires. Then I and she had together to redo the code again, it was a little bit of conflict when we tried to work together due to the different approach of the problem. To solve this, I listened and tried to integrate our ideas to make the final result correctly. Furthermore, I was frustrated with my costing part in the report due to others team members have other assignments to do and it was pending for 2 days because my part depends heavily on their works. To solve this, we hold a quick conversation on Facebook Messenger to figure things out.

Week 4: Finalising

In this final week, we cross-check again each other work and code. Tommy was in charge of consolidation all the works done by the group. In the final time, I was in charge of testing the Proof of Concept 1 which was the streaming data. Thanks to the pilot coding in the previous week, I was able to fix the bus of the simulation of streaming data made by team members. I also cleaned the code and make sure it runs smoothly on the BigVM. The problem this week was our report's word limit, it was 3395 which is over the limit of 95 words. We needed to prune our parts in the report, we had read each part of the report together and decided which part to prune The work this week runs smoothly, our team was on the track of doing the project.

Conclusion

In general, the project was smooth even though we had some problems with analysing the business case and coding task. Learnt a lot of teamwork skills and experience from my teammates who from a different background. If I was asked to do this project again, I will start to change my role to the business analyst as I have more background in it and I need to enhance my writing skills of the documentation as it has many unorganised parts that were fixed by the teammates. Furthermore, all team members are cooperative, helpful and diligent.

Wai Chun Lam (Tommy) - 28422635

Role: Coordinator, Code tester

Responsibilities:

Coordinator: This project expects different part of roles, starting from designing the direction of the whole project, discussing the main ideas for each of the parts, and combining as well as refining the ideas together. As one of the coordinator in the groups, I helped to organize the work flows and make sure all the tasks were on the right track smoothly.

Code tester: Part 3 of the project, Proof of concept required the application of Scala programming in spark-shell. Since some of my teammates were working on the code building part, my role was to assist them on checking the completeness and successfulness of the codes, specifically PoC02. I was helping to check whether the accuracy of the output and the completeness of the coding and commenting.

Diary:

Week 1–2: 30th Sept – 13th Oct

In the first two weeks, we were building up and clarifying the basic concepts through some online materials and references. Combining with knowledge throughout this unit, we have started to generate some basic ideas regarding the business project.
Working as a team, we helped each other when we had some unclear concepts on big data processing or hurdles on programming. This helped us a lot to build up a strong background to finish the tasks of the projects in the later days.

Week 3–4: 14th Oct – 20th Oct

In these two weeks, we were in a tough period with deadlines from different units. However, we still able to keep the project moving, by making good use of time even the rest time on laboratory after we achieved all the laboratory tasks.
In terms of the project, in the third week we started concluding and finalising the ideas and thoughts we had from first two weeks. Hence, we divided the project into sections for each of the teammates to work on.
In the final week, we consolidated all the progress we up to. By giving comments and feedbacks, we further fine-tuned our project and finalised all the progress in the last two meetings.

Contribution:
I involved since the brainstorming stage of the project, sharing ideas and thoughts regarding the project scope. Same as my teammates, I involved in part of the proposal design, namely the design analyses part.

Apart from the roles described above, I also helped to consolidate all different parts of the project into a final report, making sure all the contents, ideas, styles, and referencing are totally aligned throughout the whole project.

Learnt from this group work:

Throughout the projects, I have acquired the knowledge on how to set-up a business project on big data analysis, especially on the analyses designing section where I mainly focused on. With sharing the idea of the other sections with my teammate, I have also learnt how to budget a big data analysis project. (i.e. identifying the size of data, specification of hardware required, etc.)

How have you learnt? (i.e. learning techniques)

As this project involves knowledge from different aspect of big data processing, we spent some time to go through some basic materials and then divide the project into sections for each teammate to focus deeply separately at the beginning. We then collaborated and shared the knowledge and ideas we got.

What went well about this project?

Throughout the whole project, although we had a tough timeline including other units to cope with, we worked closely and efficiently in and between every meeting. One of the key made us finished the project smoothly was keeping reviewing and sharing thoughts on our findings.

What went wrong?

For the coding part of POC01, we were able to run the k-mean clustering on streaming data. However, we were predicting the data all in same clusters, which is not we expected. Therefore, we have gone through the documentations and finally fixed the errors successfully.

Over conclusion:

During the project, I have learnt how to design a business project on big data processing. With the strong collaborations with my teammates, we are able to achieve the knowledge and complete the tasks successfully. If I was asked to do the project again, I would put more focus on the coding part. It is because some of the parts are taking longer time than we expected.

Wongnaret Khantuwan (Rex) - 29289440

Role: Leader and Programmer

Responsibilities:

Main responsibility

- Write section Design Analyses part c-d-e, Costing
- Implement, draw diagram and document POC1
- Assign tasks to other team members

Reflective diary

Week 1: 30th September to 6th October

I started the project by a meeting in our tutorial. We mostly focused on understanding the requirement for this project. I allocated the task for all members and was responsible for clarifying all requirements. I felt quite confident about this project as our team did well so far. In short, I learnt how to organise and allocate the task in a team project like this.

The good thing is that our tutor explained most of the unclear point for this assessment. There is no problem for this week.

Week 2: 7th October to 13th October

We focused on the first two sections of the report, which is the business case and business requirement. I helped Rita to find more sources and idea for the business case. Also, I found the price of different types of servers and machines and integrate it with Thomas. I learnt a lot about the advantages and disadvantages of each kind of servers. I was delighted because everyone understood their part, and our team was on the right progress.

The plus point is that we already found a way to deal with the business case and user. On the other hand, we started having a problem concerning matching the availability together. Finally, we agreed that Thomas and Rita would go to the consultation next week to confirm all our doubts.

Week 3: 14th October to 20th October

As all the team members were busy with their assignments, we decided to cancel the meeting this week. In steading of the meeting, we updated our work via Facebook messenger, Google Drive and Trello. Thomas and I kept researching about design analysis section and costing section. Besides, I also rearrange the task and deadline to fit with each member's availability. I learnt how to make a backup plan for a project and how to react to an unexpected situation.

Although we have not enough time to meet face to face, everything still went well. Rita has uploaded her work, Tommy and Thomas also had significant progress. The minus point is that we did not have enough time to double check the work from other members this week. To deal with it, I set another deadline, and we will spend more time testing next week.

Week 4: 21th October to 27th October

This week, as most the team members were released from their assignment, we spent more time on this project. I completed my design analysis and costing my final version of the business requirement and business case. Apart from that, I created a simulated data and started trying with the POC1. I was pleased because this was a productive week, and most of the major parts were finished.

On the bright side, we are confident that we can meet the deadline for this project. The only problem is that the POC1 section still got some bugs. We were expected to solve the bugs in the next few days by reading the documentation of the library.

Final week

In this final week, we cross-checked each other work and code again. I wrote my very last part of designing part of the report and finishing code for Proof of Concept (PoC) 1 of streaming data. For the last two days, I created the testing and training data for streaming to test the k-means algorithm. I feel a little bit hard when coding the K-means streaming algorithm because the documentation of this library of mllib is not completed. At first, we got a wrong prediction result. To resolve this, I have my team to help with the coding part and the documentation, so I will have time to fix the bugs. Finally, we realised that the main problem came from the unclean dataset, so we cleaned the dataset again, and everything worked well. I was so happy. Our group had completed all specs, and we met the deadline set in the beginning.

Conclusion

Through this project. I have enhanced my leadership skills by leading the group to perform different tasks on this big assignment. I have combined the capabilities of my team even though our team is formed from people with different backgrounds to do the project to run smoothly. If I were asked to lead the team for this project again, I would keep my team updated on the work I am doing, and sometimes I have miscommunications with my team.